

## **Best Practices –Preparation of Address Data for Geocoding**

### **File Types**

Your input file can be several different types. ArcView can geocode Access tables, Excel or dBase (DBF) file types. Delimited text (TXT) will also work, but only as an intermediate file type. (It will need to be imported into Access. )

#### **Access**

Access table as input for geocoding is the recommended format. It is the less likely to have extraneous formatting or margin data compared to Excel.

#### **Excel**

There are some specific formatting requirements to use Excel files with GIS. If you have multiple worksheets, process one worksheet at a time.

#### **DBF files**

In Microsoft Office 2007 it is not possible to convert a file to DBF file format. However you might have a DBF file created from some other application. DBF files can be geocoded without converting to another file format.

#### **Delimited ASCII (text)**

If you have delimited text, it is because you are likely extracting this data from some application that generates this format. If you have an option at all for a field delimiter, make sure that you choose something besides the comma (.). Address data often contains commas, and the data will be shifted over into the next column. The recommended alternative is the pipe (|) symbol, generally found above the backslash (\) key.

#### **A note about Office 2007**

ArcView doesn't support either Access or Excel 2007. Save the file back to the 97-2003 format (Save As from the Microsoft Office icon.)

#### **A note about data that was copied from a web page**

Data that was copied from a web page into Excel should always first be saved in the TXT file format. This will remove any formatting (HTML) that is still associated with the data. Then you can save it back into a new Excel worksheet or import it to Access.

## Address Format

These address format guidelines are the same for all file formats. Details specific to a particular file format follow.

### Separate columns for address components

Satellite-clinic Table.xls							
	A	B	C	D	E	F	G
1	Agency Number	Satellite	Satellite Clinic	Address	City	State	Zip
2	60300		Platte County Health Department	212 Marshall	Platte City	MO	64079
3	10500		Hamilton City Hall	200 S. Davis	Hamilton	MO	
4	10500		Polo Christian Church	301 Main St.	Polo	MO	
5	11200		United Methodist Church	312 E. 7th	Mound City	MO	64470
6	51400		Noel Housing	624 Johnson Dr.	Noel	MO	64854
7	51400		Tyson Chicken Plant	Hwy 59	Noel	MO	64854
8	51400		High School WIC	110 Mustang Dr.	Anderson	MO	64831
9	60500		Lawson WIC Clinic	D Highway	Lawson	MO	64062

Address components should be divided with separate columns for address (i.e. 123 N Main St), city, state and ZIP code. Data extracted from some applications have separate columns for house number, directional prefix, street name and street type. These must be concatenated into a single address column for geocoding.

- The input file should contain at a minimum: **Address, City, State, Zip**
- All other fields, such as patient ID, county, facility name, comments etc can remain in the input table. This information will be retained in the output table.
- Each address component should be given a simple column name. Preferred names: Address, City, State, ZIP. Alternately, if multi-part addresses are used, Address1, Address2, City, State, ZIP.

### Multi-part addresses (primary and secondary addresses, suites, apartments etc)

Additional information beyond a basic street address (e.g. PO Boxes, buildings, suites, apartments), must come after the main part of the address. (Right: 123 N Main St, 1<sup>st</sup> floor, Wrong: 1<sup>st</sup> floor, 123 N Main St). In the 2<sup>nd</sup> example, the software tries to interpret 1<sup>st</sup> floor as 1 Floor St, and never gets to the real part of the address.

If you want to retain this additional information, create a second column and transfer this extra information into a secondary address column. Depending on what sort of information it is, the geocoder might not use this information. You just need to make sure that all the basic address (123 Main St) is in one column.

If the address is only a PO Box, Rural Route or similar style, leave this information in the first address column. You will only be able to get to the closest ZIP code with this type of address.

Tip: To help sort out addresses that appear in both an Address1 and Address 2 field, sort each address column in descending order. The addresses that start with a letter (PO BOX, Rural Route) will be sorted first, followed by any column starting with a number.

### No # sign

House numbers or one or two digits are sometimes preceded by a # (pound) sign. (Example: #10 Main St). These symbols need to be deleted.

*Tip: If you sort your address column, these will appear at the first of the listing, before any of the numbers.*

### **Directional prefixes**

Use directional prefixes whenever one is available. When a city or town uses directional prefix (N S E W), and a prefix isn't provided in the input address, possible duplicate locations may be found. (e.g. 123 N Main St, 123 S Main St are the same if you exclude the direction).

### **Extraneous data within the address field**

If there are additional comments within the address field beyond supplementary address information (the use of the word 'unknown', question marks (??), or any other non-address related information) these comments should be removed. They do not contribute information to the process and generally guarantee the record will fail to match. If you must retain this information, create a new column, and transfer the information to it.

### **Street Names w/ abbreviations**

Street names containing abbreviations will frequently fail to find a match. You can spell it poorly, but you have to try to spell it out. Some common ones:

Martin L King

ML King

MLK

JFK

James CP Bell (this stands for James Cool Papa Bell)

MO Blvd (all state abbreviations must be spelled out – in this example change to Missouri Blvd).

*Note: Some French or Spanish names may require a space within the name to match correctly.*

*Examples: Right: De Soto, De Paul, Wrong: Desoto, Depaul.*

### **Numbered Streets**

Where streets are named with numbers instead of names, the numbers should always be listed as a number. Frequently numbers from 1 – 10 are spelled out (First vs. 1st). These may be returned as unmatched (depends on the geocoding software you are using). Change these names to their numeric equivalent (Right: 1<sup>st</sup>, 2<sup>nd</sup>, Wrong: First, Second).

### **Addresses based on highways**

Addresses on highways, whether they are US highways, county routes or even frontage roads are problematic for two reasons.

- 1) The source road network used by most common geocoding software doesn't include many house numbers for highways. The highways appear with their name, but the house numbers are all 00000 (zeros).
- 2) The highway name has to be very specific. For example, if the address is 1234 C Hwy, the geocoder may have it as 1234 County Route C. The exact rules on the naming convention for each 'level' of highway still need to be researched. The correct name is based on the US postal service.

- 3) Counties which have their address style like 1234 Lawrence 2457 won't process correctly. Try changing these to 1234 County Rd 2457.

### **City Names – St. Louis City**

St. Louis City is not a city name; it is a county name. St. Louis County is also not a city name. For geocoding, make sure these names are edited in the city name columns to St. Louis.

Standard abbreviations for Saint and Mount are allowed (can be St and Mt). All other city names should be spelled out. (Right: Kansas City, St. Louis, Springfield, and El Dorado Springs. Wrong: KC, STL, Spfd, and El Dorado Spgs)

### **State**

All input data should include a column for STATE. It is common for users to leave this column out, assuming everyone will know the state is Missouri. (We actually routinely deal with data out of state.) The software used for geocoding can't make these assumptions. If there is no state column present in your data, you must add one, and populate it with a state abbreviation (MO).

### **ZIP code**

ZIP codes should be stored as text field types. Some ZIP codes are stored as ZIP +4. Storing the data in a text field will accommodate the '-' used in this style of ZIP code.

- a web site, such as <http://zip4.usps.com/zip4/citytown.jsp> to help identify missing ZIP codes. If the rest of the address is correct, except for a missing ZIP, you will likely still get a match, but Best Practice is to have data in every field.

## File Preparation – Excel

Here are some general guidelines for making any Excel file ‘geocode ready’ . If your Excel file meets these criteria, you can use it in ArcView without translating it to another file format.

- Keep a copy of your original file.
- Use only one line for your column headings, don’t create ‘stacked’ headings
- The first line must contain ONLY the names of the fields; brief headings are preferable.
- There should be NO additional header, footer or sidebar comments in the document. This includes summary lines, totals, references, citations, etc.
- Headings should start with a letter, not a number.  
Example: 2002rate should be changed to rate02 or rate2002
- Brief headings are preferable.
- Field names should be in all UPPER case or all mixed/proper; either is OK, as long as it is consistent.
- The same consistency applies within the data. Text fields should contain data in all UPPER case (e.g. COLE, OSAGE, BOONE) or all mixed/proper case (e.g. Cole, Osage, Boone).
- For number data do not use 1000 separators. This can be changed under Format Cells > Number Uncheck ‘Use 1000 separator (,)’.
- As long as you have all the required address fields, any other data you want to be associated with the output shape file can be included; you aren’t limited to just the address data for geocoding.
- Remain consistent with your capitalization. If you have multiples files, they should contain data in all UPPER case (e.g. COLE, OSAGE, BOONE) or all mixed/proper case (e.g. Cole, Osage, Boone).
- None of the cells should be a calculated or summary value based on the data of other columns. Create a new column and Copy and Paste Special, selecting Values only as you paste to convert to plain numbers (or text).
- Within the data itself, particularly for character/text data make sure there are no leading blanks. Look closely at how the data aligns in a column. Text should appear justified to the left, without spaces before the first character.
- **Bold** text, boxes around cells and other cosmetics can be left as-is. This formatting will be stripped off automatically during the conversion from Excel to the other formats.
- Don’t highlight cells or otherwise color code them to somehow classify your data; this information will not be available to you in ArcView. Instead, add additional columns and create a value to represent the meaning of the colors.

## File Preparation – Access

Here are some guidelines for making your Access table ‘geocode ready’ Several are common to all file formats:

- Keep a copy of your original file.
- Tables in Access should have a field defined as an AutoNumber (suggested name ID) and set as the primary key. If you are importing a file into Access, this can be done as part of the import process.
- Don’t use punctuation characters (- & % # / etc) in your column names. There is one exception; you can use \_ (underscore).
- Headings should start with a letter, not a number.  
Example: 2002rate should be changed to rate02 or rate2002
- Field names should be in all UPPER case or all mixed/proper; either is OK, as long as it is consistent. . (This won’t really cause you problems; it’s just a Best Practice).
- Remain consistent with your capitalization. If you have multiples tables, they should contain data in all UPPER case (e.g. COLE, OSAGE, BOONE) or all mixed/proper case (e.g. Cole, Osage, Boone).
- Within the data itself, particularly for character/text data make sure there are no leading blanks. Look closely at how the data aligns in a column. Text should appear justified to the left, without spaces before the first character.
- As long as you have all the required address fields, any other data you want to be associated with the output shape file can be included; you aren’t limited to just the address data for geocoding.
- 
- There are several field names that are reserved and should not be used in Access. They can be found at: <http://support.microsoft.com/default.aspx?scid=kb;EN-US;q248738>

Access tables do not need to be converted to any other file type.

*Last Updated December 10, 2008*

*Tracy Schloss*